

# Beyond Human: Deep Learning, Explainability and Representation

M. Beatrice Fazi 

University of Sussex

Theory, Culture & Society

2021, Vol. 38(7-8) 55–77

© The Author(s) 2020



Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/0263276420966386

[journals.sagepub.com/home/tcs](https://journals.sagepub.com/home/tcs)



## Abstract

This article addresses computational procedures that are no longer constrained by human modes of representation and considers how these procedures could be philosophically understood in terms of ‘algorithmic thought’. Research in deep learning is its case study. This artificial intelligence (AI) technique operates in computational ways that are often opaque. Such a black-box character demands rethinking the abstractive operations of deep learning. The article does so by entering debates about explainability in AI and assessing how technoscience and technoculture tackle the possibility to ‘re-present’ the algorithmic procedures of feature extraction and feature learning to the human mind. The article thus mobilises the notion of incommensurability (originally developed in the philosophy of science) to address explainability as a communicational and representational issue, which challenges phenomenological and existential modes of comparison between human and algorithmic ‘thinking’ operations.

## Keywords

algorithmic thought, deep neural networks, explanation, incommensurability, interpretability, philosophy, XAI

## Beyond Human Representation

The success of the Google-owned artificial intelligence (AI) company DeepMind and its computer program AlphaGo is well known. In March 2016, AlphaGo defeated the 18-time world champion Lee Sedol at Go, an ancient, complex game that involves moving black and white stones on a board to control territory. The victory was widely reported by news outlets, with commentators drawing parallels with the famous 1997 chess match between the grandmaster Garry Kasparov and the IBM supercomputer Deep Blue. The performance of DeepMind’s AlphaGo, however, is considered more striking than that of its IBM

---

**Corresponding author:** M. Beatrice Fazi. Email: [b.fazi@sussex.ac.uk](mailto:b.fazi@sussex.ac.uk)

**TCS Online Forum:** <https://www.theoryculturesociety.org/>

predecessor, as the board game Go involves many more possible moves than chess. Go also requires strategic skills that are more intuitive than those useful in a chess game, and which are therefore less mechanisable.<sup>1</sup> In this respect, AlphaGo's victory also made headlines because it demonstrated the potential of the type of AI technology that DeepMind champions: *machine learning*. This expression denotes computing techniques that provide computer programs with the ability to improve over time with minimal human intervention when exposed to large amounts of data. The same programs subsequently apply this 'learning' to make data-driven decisions.

While AlphaGo's success is well known, the story of DeepMind's 2017 cognate program AlphaGo Zero is less familiar to the general public. AlphaGo learned to play Go by being exposed to training data derived from millions of moves of past players. AlphaGo Zero, in contrast, was not given data from games played by humans or machines but was trained by playing against itself, starting from random moves and knowing nothing about the game of Go. This feat of solid and stable *reinforcement learning* amazed the AI community, which welcomed AlphaGo Zero as a significant achievement.<sup>2</sup> DeepMind proposed a self-taught AI program that can train itself from scratch, being de facto 'no longer constrained by the limits of human knowledge', as DeepMind put it (see Hassabis and Silver, 2017). Whereas AlphaGo took months to learn how to play, AlphaGo Zero took just a few days, less computing power and a streamlined architecture to master the game, quickly reaching levels of 'superhuman performance' (Silver et al., 2017: 354).

The ability of a program to self-train without the input of human data is a key step towards achieving the holy grail of AI research: *artificial general intelligence*. This is the capacity of a machine to perform a breadth of cognitive tasks like that of a person. Recognising this, DeepMind is keen 'to make some real progress on some real problems' (Hassabis, quoted in Gibney, 2017) and extend its success to areas with practical applications (e.g. material design, genomics and drug discovery). Central to this possibility is acknowledging that being 'no longer constrained by the limits of human knowledge' means that AlphaGo Zero won not by out-reading humans but 'by seeing patterns and shapes more deeply', as Andy Okun – the president of the American Go Association – observed (quoted in Sample, 2017). In other words, AlphaGo Zero succeeded not because it behaved like a human player but because it played differently from a human. This condition is particularly interesting from both philosophical and sociocultural perspectives: in my view, cases such as AlphaGo Zero allow us to say that contemporary developments in cognitive computing are departing from what, in a previous work, I called the *simulative paradigm*, which has been looming over AI research since Alan Turing's proposition of an 'imitation game' (Turing, 1950) to test the cognitive capabilities of an artificial system.

In that work, I claimed that we should conceive of ‘automated modes of thought in such a way as to supersede the hope that machines might replicate human cognitive faculties, and to thereby acknowledge a form of onto-epistemological autonomy in automated “thinking” processes’ (Fazi, 2019a: 813). I thus argued for the possibility of considering the algorithmic modes of thought of computing machines as ‘dramatically alien to human thought’ (Fazi, 2019a: 813). This article continues developing that line of argumentation and focuses on algorithmic modes of cognition, thus maintaining a philosophical commitment to ontological and epistemological questions about the nature of thinking in the 21st century. More specifically, I here consider *algorithmic thought* (what it might be and might do) by engaging with some theoretical implications of a computer program being ‘no longer constrained by the limits of human knowledge’, as DeepMind claimed AlphaGo Zero to be. I do not simply repeat what the AI industry says about itself and its products, however; I instead critically address these claims to assess their philosophical consequences by interpreting this declared freedom from human knowledge as a form of autonomy from human modes of *abstraction* and by relating these issues to questions about *representation*.<sup>3</sup>

This article thus continues to develop my theorisation of algorithmic thought by addressing the contemporary expansion of automated modes of abstraction that operate via what computer science calls *representation learning*. As the computer scientist Yoshua Bengio has put it, the central principle of machine-learning methodologies is ‘the automated discovery of abstraction’ (2013: 3). ‘Representation learning’, LeCun, Bengio and Hinton explain, ‘is a set of methods that allows a machine to be fed with raw data to automatically discover the representations needed for detection or classification’ (2015: 436). In this article, I focus on precisely this aspect of current developments in AI technologies: how the extraction and organisation of ‘discriminative information from the data’ (Bengio, 2013: 2) that these technologies perform is specific to their computational character and how it consequently transcends or is independent of human access. I thus consider the 21st-century development of computational procedures for which, at present, no adequate human cognitive representations exist and for which, significantly, human cognitive representations are also unnecessary.

## **Black Boxes of Decision-Making**

Research in machine learning is at the forefront of the agenda of AI and data science. As an umbrella term, ‘machine learning’ denotes not a single computational technique but a plethora of often quite different tools and approaches to cognitive computing. These approaches have been grouped together under this label because they all involve algorithms that can ‘learn from experience’ insofar as they can change their

operations to better fit the requirements of their tasks. These requirements are specified in the data that these algorithms must handle. ‘Machine learning automates automation itself’ (Domingos, 2015: 9–10), for ‘computers can learn programs that people can’t write’ (Domingos, 2015: 6). Machine learning thus involves ‘a change in programming practice’ as well as in ‘the programmability of machines’ (Mackenzie, 2018: 21). This condition has been described as a ‘quiet revolution’ (Alpaydin, 2016: ix), as a new season after a long and harsh winter in AI research, and as a computational renaissance precipitated by a novel form of AI that in fact draws from old cybernetic ideas.<sup>4</sup>

Since there are many machine-learning techniques and many devices that successfully implement them, it is important to clarify that *deep learning* is the approach followed by both AlphaGo and AlphaGo Zero, and the method that Google’s DeepMind echoes in its own name. Deep learning is itself a remarkably multifaceted technique. To simplify, an artificial neural-network system relies on layers of artificial neurons to process information. These layers of artificial neurons are connected and influence each other in a complex web of interacting units, somewhat like biological neurons are understood to do in a biological brain. A lower layer of neurons performs a computation and transmits this result to the layer above, enriching the final outcome of the layer at the top. What is obtained in each layer is a new representation, ‘which can be used as input for deeper layers’ (Bengio, 2013: 4). A neural network is said to learn, then, because it can tweak its calculations and modify its interactions by tuning parameters via activation and back-propagation among layers until the desired output (i.e. the desired final representation) is produced. The network, however, is called *deep* if its structure encompasses intermediary ‘hidden’ layers between the input and the output.<sup>5</sup> The architecture of a deep-learning system differs from that of a standard artificial neural network precisely because of the presence of these multiple non-linear hidden layers.

Deep techniques are often discussed, as they promise to accelerate the computational automation of today and fuel the digital transformations of tomorrow. Although artificial neural networks have been around for decades (they are a core technology of *connectionism*, a biologically inspired approach to AI that emerged in the 1980s), it is only in the past decade, thanks to the volume, velocity and variety (see Beyer and Laney, 2012) of Big Data and the increase in computational power, that AI research and industry have begun to capitalise on artificial neural networks’ potential. Computational problems that the AI community thought could not be tackled for many years – such as recognising speech and other intricate patterns in high-dimensional data – are now being significantly improved. Beyond reports on its promising results (and the hype about its achievements) that have appeared both in

specialised literature and the mainstream media, deep learning has also caught popular attention in a less flattering light. Because of how a deep neural network operates, relying on hidden neural layers sandwiched between the first layer of neurons (the input layer) and the last layer (the output layer), deep-learning techniques are often opaque or illegible even to the programmers that originally set them up. While ‘different machine learning models provide different levels of interpretability with regard to how they reach a specific decision’, it is thus commented that deep neural networks ‘are possibly the least interpretable’ (Kelleher, 2019: 245). In this sense, deep-learning programs are said to be *black boxes*: it is clear that they work but often is not equally clear how or why.

In computing and engineering, the expression ‘black box’ is borrowed from cybernetics and used to describe an object or a system that is viewed uniquely in terms of its inputs and outputs, and whose internal working therefore remains concealed. When approaching a black box, one is interested only in stimuli and responses; one considers what goes in and what comes out of the black box, not its inner components or operations. While some computer and data scientists might take issue with the popular claim that deep-learning systems are black boxes,<sup>6</sup> there remains the fact that, once a deep neural network is trained (or self-trained, as in the case of AlphaGo Zero), it can be extremely difficult to explain why it gives a particular response to some data inputs and how a result has been calculated. The strength of a deep neural network lies in its capacity to find non-linear patterns in large datasets and improve this extraction through iterative interactions. The other side of the coin, however, is that the automated learning choices of a deep neural network are not yet fully understood by programmers. The knowledge generated in these models remains, in part, implicit due to the non-linear nature of deep learning, its compressed information, and the distributed character of the network’s representations, which rely on the many configurations of its large sets of variables. Such a complex, layered architecture entails difficulty in analytically comprehending what nodes and layers have learned and how they have interacted to transform a representation at one level into another representation at a higher, more abstract step. Moreover, interpretability is not a standard feature of deep learning also because of the difficulty of producing a satisfactory mathematical theory as a foundation for these architectures. Interestingly, what makes deep techniques powerful also often makes their theoretical underpinning tentative. Progress comprehending these computational activities is achieved by trial and error, and operations are often rationalised retrospectively. To put this otherwise, ‘many algorithms using artificial neural networks are understood only at a heuristic level, where [scientists] empirically know that certain training protocols employing large data sets will result in excellent performance’ (Lin et al., 2017: 1223).<sup>7</sup>

The trope of the black box is a recurring one in the sociology of science and in science and technology studies. Famously, Bruno Latour (1987) described parts of science that have been accepted and are no longer controversial as black boxes. Science, for Latour, can become a black box when its inner workings are no longer open for scrutiny or debate, when consensus has been reached about certain results, when the success of a theory or a method obscures how scientific and technical work operates, and when a hypothesis is settled as a matter of fact. So, paradoxically, ‘the more science and technology succeed, the more opaque and obscure they become’ (Latour, 1999: 304). Always within the diverse domain of the sociology of knowledge, social constructivist positions – such as the *social construction of technology* (or SCOT) – stress the need to ‘open the black box’ to recognise the interpretive flexibility of an artefact while also crediting ‘users as agents of technological change’ (Kline and Pinch, 1999: 113). Opening the black box thus involves counteracting the closure mechanisms that ‘play a part in bringing about both scientific agreement and the stabilization of an artefact’ (Pinch and Bijker, 1984: 425).

Considering deep neural networks, concerns about AI as a black-box technology in part recall these earlier debates in science studies yet also transcend them. The black-box character of deep-learning techniques is, first of all, a *technical condition*. Of course, these techniques are part of the contemporary world, and their predictions, classification and clustering impact the everyday lives of millions of people; with respect to this impact, a social constructivist perspective proves useful to explain the concurrent yet multidirectional involvements of relevant social groups and the values and interests that inform their participation. However, if opening the black box means asking ‘how technology is made’ – to paraphrase the title of a famous essay in social constructivist technology studies by Bijker (2010) – then, while doing so, we cannot avoid addressing the ontological and epistemological specificities of that same technology.<sup>8</sup> In the case of deep learning (and machine learning, more generally), my proposition is that we should not overlook the computational and increasingly autonomous character of these technologies.

Insofar as they are computational, these are calculative techniques that involve quantifying and systematising the real world through discrete functions. Above all, since they are computational, deep-learning methods involve *decision-making*. In a computational context, decision-making is the mechanised process that results in the selection of a particular result or output among possible alternatives. This decisional capacity of computational systems is inscribed in the definition of an algorithmic procedure: a step-by-step ‘effective’ method to address a problem that can be posed as a yes-or-no question of input values (Turing, 1936). In addition, however, it is crucial to consider how deep learning algorithms are also increasingly autonomous artificial actors

(see Fazi, 2019b), requiring little engineering by hand. In these *learned* (not designed) systems, transparency is a matter of accountability vis-à-vis their automated and quasi-autonomous decision-making capacities, which have been transferred from humans to the AI system. Thus, while all technologies are black boxes to an extent (even a door handle can be approached as one because knowledge of its inner working is not necessary to open a door), the consequences of the black-box character of deep learning are different because, in this case, it is agency itself (that is, the capacity of the technological system to operate upon its environment) that is opaque.<sup>9</sup>

The decision-making of quasi-autonomous artificial agents powered by deep learning affects millions of people every day. The range of decisions covered by deep learning is vast. It pertains to mechanisms of classification, clustering, ranking and pattern-finding, which are employed, for instance, in credit card fraud detection, spam filters, search engines, market segmentation, social media advertising, insurance and credit scoring, healthcare management, transport and logistics, loan qualification and mobile communication. These and other operations were often determined by humans in the past; today, the human user rarely has a concrete sense of the reason or mechanism of certain results or what inputs they follow. The task left to the social scientist, cultural theorist, philosopher, legal scholar and critical theorist is asking what would count as ‘cracking open’ these AI black boxes responsible for so much contemporary decision-making – particularly now that society has entered an era of computational applications whose success is measured by the capacity of computational agents to act on their own. Although involving various arenas of public and academic discussion, this question has been most explicitly developed within the interdisciplinary scholarly debate about the politics and governance of algorithms (see, for instance, Amoore, 2020; Ananny and Crawford, 2018; Beer, 2018; Benjamin, 2019; Noble, 2018; O’Neil, 2016; Pasquale, 2015). It is impossible to review these rich discussions in full here: suffice it to say, however, that there is consensus on the fact that automated cognitive agents processing increasingly vaster amounts of data will play ever more significant roles in regulating and directing our lives. What academia and the general public alike are asking for is transparency regarding how security, government, media, retail, finance, science and industry employ AI on a daily basis, often to influence human action. *Explainability* is a key word for present and future algorithmic cultures, raising equally unique social and ethical challenges.

## **A Different Kind of Abstraction**

From the perspective of this article’s engagement with computational operations supposedly beyond human knowledge, both the notion of

and debate about *explainable AI* (XAI) are significant and relevant, as they call for the opaque powers of AI to be leashed in the realm of observation so that the mysteries of machine learning eventually surface. I am not referring here to the visual form of machine learning (i.e. to how data and data patterns can be made visible thanks to specialised graphics showing something about how a machine-learning algorithm's output relates to its inputs – see, for instance, how this is discussed in Mackenzie, 2015). Rather, I am gesturing towards the more figurative sense in which technoscience and technoculture are addressing the possibility to present (or *re-present*) algorithmic operations to the human mind and thus make the abstractive operations of artificial cognitive agents (and their internal representations) somehow available to and comprehensible within the epistemic landscape of human cognitive representation.

Interestingly, in her sociocultural study of machine-learning algorithms, the information scholar Jenna Burrell distinguished between three types of opacity: (1) opacity as corporate or state secrecy (i.e. algorithms as proprietary, their lack of transparency a form of institutional protection to maintain trade secrets and competitive advantage); (2) opacity as technical illiteracy (i.e. writing and reading code as highly technical skills that require specialised knowledge and are thus inaccessible to most people); and (3) opacity as an inherent characteristic of machine-learning techniques – that is, 'opacity that stems from the mismatch between mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of semantic interpretation' (2016: 2). The debate on explainability in AI concerns all three types of opacity, but the last one, which pertains to specific techniques used in machine learning, is the most conceptually challenging. This is a form of opacity that, in the case of deep-learning systems, thrives upon the complexity of their high-dimensional domains – a complexity for which a machine might build a model but a human most likely could not hand-engineer one. Technically speaking, the crux of the problem of explainability in deep learning lies in artificial neural networks not returning clear representations of their inner workings to programmers. Deep neural networks lack model interpretability, so when considering why a machine made a particular decision or one prediction instead of another, we remain ignorant at worst and agnostic at best. Returning, for instance, to the case of DeepMind and its AlphaGo machines, to understand how and why AlphaGo or AlphaGo Zero chose a particular move instead of another, the justification given by the program may consist of a rendition of the network's weighted connections and how these pass their outcomes to the next layer in the neural network. Of course, this might not signify much to a human user, for the calculation the neural network carries out cannot be easily performed by a human mind. Even if this performance were possible, however, it might



be objected that merely rendering the calculation would hardly count as a meaningful form of understanding.<sup>10</sup>

This issue links to another open question about when and why an explanation might be considered useful and successful or not. I discussed earlier how establishing a theoretical ground for deep learning could help programmers interpret the choices a deep neural network makes and thus validate its behaviour. Here, however, it should be added that, alongside explorations of deep learning's mathematical foundations, the growing field of XAI focuses on the taxonomies of *desiderata* and of methods for interpreting AI systems. Research in XAI often explicitly looks for pragmatic approaches to human-readable explanations that can meet the expectations of end-users, whether they are medical doctors and patients in an automated diagnosis scenario or banks and their customers agreeing on a mortgage assessment. Questions about the nature and characteristics of a successful explanation are thus also answered by considering the social dimension of interpretability, and they must confront the fact that, when attempting to produce knowledge about a deep-learning system's input-output relationship and the aggregate behaviour of its decision structure, 'we may not even have the words to express the concepts that some parts of the model represent' (Spreeuwenberg, 2019: 32).

In this respect, deep learning may be changing the epistemic possibilities of justification and explanation, effectively reshaping how science imparts information and knowledge. My claim, however, is that deep learning is changing the meaning, scope and use of abstractions as well. To expand on this point, it is useful to distinguish between the *modus operandi* of the traditional statistics community and the machine-learning community, which the statistician Leo Breiman (2001) elaborates on in a much-cited paper. Breiman speaks of 'two cultures' to explain this distinction: on the one hand, traditional statistics assumes that data models are the best way to solve problems; on the other, scientists working with machine learning believe that algorithmic models can do better. Breiman's paper attempted to show that the data models of statistics are not applicable to a wide range of problems, so statisticians should allow a wider variety of tools to be employed in their discipline. Breiman himself is a pioneering scholar who helped bridge the gap between computer science and statistics, writing and working when machine-learning techniques were still underexplored in statistical science.

In what follows, rather than lingering on Breiman's advocacy for machine learning, I focus on how his paper addressed black boxes. From a scientific perspective, Breiman argued, nature is a black box: the challenge is to extract *information* on how nature associates input variables to output variables and to produce *predictions* about these input-output relations. Data are what scientists handle to consider precisely these tasks, and models are what scientists use to draw conclusions

from data to produce these descriptions. The relationship between data and models, however, is different in statistics and computer science. Breiman explained that, traditionally, the purpose of statistics is to produce an understandable picture of the relationship between the input variables and the end results in the phenomenon or situation observed. Of course, nature is overwhelmingly complex and rich with variables; statistics hopes to achieve, at best, accurate representational approximations, in which a data model is as close as possible to represent, and thus explain, the black boxes of nature. The development of algorithmic methods of statistical analysis, however, involved doing things differently. The black box of nature remains an unknown whose underlying workings are not the target of scientific enquiry. The aim of algorithmic models is not to find the 'true' data-generating mechanism but to use an algorithm to account for that mechanism as well as possible. In this sense, computer science is less concerned with explanation than with predictive accuracy, and modelling is treated as a problem of function optimisation. 'The goal is not interpretability, but accurate information' (Breiman, 2001: 201).

Breiman's argument, of course, is not the only reconstruction of the field of machine-learning research. However, drawing from Breiman's account, one can begin to explain how and why, thanks to the contemporary availability of high computing power and of vast amounts of data, previously undetected or underrated differences between explanation and prediction have moved to the fore of scientific practices, such as statistics.<sup>11</sup> Moreover, the difference between explanation and prediction highlighted in Breiman's paper also speaks about what abstraction is – or can become – in algorithmic modelling. Following Breiman, we could say that statisticians want interpretable approximations of what they hypothesise happening in the elusive black boxes of nature. In this sense, they use data to *abstract away* a model.<sup>12</sup> In fact, according to Breiman, statistics ends up focusing more on the model than on the problem or the data themselves. Pushing Breiman's comments further, we could also argue that abstractive procedures work differently in algorithmic modelling, which seems to acknowledge that human abstraction might never be a particularly accurate predictor. Rather than description, then, *construction* is the epistemic tool of choice: instead of reducing a black box to fit a simpler model, the algorithmic modelling of machine learning constructs and stands as *another black box*, thus freeing abstraction from its reductionist role as a means of simplification and description. Abstractive operations of classification and generalisation have overcome the boundaries of the human mind and are performed via the weights of digital triggers in artificial neural networks. Algorithmic modelling, consequently, is not a means of interpreting but rather constructing new, complex worlds in equally new, complex computational ways. In her book on models and simulations, the philosopher Margaret Morrison observed that scientific inquiry 'involves

reconstructing or recasting nature in a specific form and investigating how it behaves or might behave under certain circumstances.’ ‘Although we can use mathematics to do this,’ Morrison continued, ‘the notion of “reconstruction” can also be instantiated in other ways’ (2015: 2). The operational black boxes of machine learning also seem to be one of these other ways, according to which epistemological reconstruction assumes a life of its own via algorithmic models that do not aim to represent and thus do not wish to explain.

## Incommensurability

Possibly due to much scientific research in deep learning focusing (quite successfully) on computer vision, metaphors or analogies that refer to the sense of sight are frequently used to describe the operations of deep neural networks. So, for instance, it is often said that these artificial cognitive agents ‘see’ visual inputs differently.<sup>13</sup> In light of what I have discussed so far, however, I claim that deep learning not only involves a distinctive type of sensibility (i.e. a different capacity to receive data inputs) but also concerns a specifically computational relation with the intelligible (i.e. with what is apprehensible only through forms of abstractive activity).

To exemplify this claim, let us consider machine learning’s increasing ability to recognise human handwriting. This is something notoriously hard to perform computationally and for which more traditional programming techniques do not work well because it is difficult to prefigure and then encode an instruction that would formally describe such a task. In other words, relatively simple, immediate human intuitions of how to identify shapes are not easily expressed in computational terms. With deep learning, however, the situation changes.<sup>14</sup> Let us assume that we want a program to recognise a handwritten digit, such as zero. In the case of *supervised learning*,<sup>15</sup> thousands of scans of handwritten zeros are fed to the machine as training examples. The program then learns to recognise the digit, not how a human might (e.g. determining that a zero resembles a vertical oval), but by mechanically detecting complex patterns of darker and lighter pixels expressed as matrices of numbers. This is arguably a different form of perception (or of input reception), and ground-breaking research on how a computational system can elaborate visual information that humans cannot even receive or perceive is being developed in the field of computer vision.<sup>16</sup> The point here, however, is that beyond physical data reception, we are also witnessing a specific form of abstractive capacity – one akin to an automated mode of conceptualisation, that is, an automated mode of forming internal representations meant to generalise while abstracting from observed facts or phenomena. In the case of deep learning, the possibility of concept formation must be understood vis-à-vis the machine’s automated

*feature extraction* from raw data. ‘Features’ are the properties and characteristics of data that the system learns to distinguish and organise in order to recognise patterns, make predictions and classify tasks: deep neural networks are algorithms for classification from features, and deep learning is largely feature learning.

None of this implies that feature learning and conceptualisation are identical. I am addressing the two operations together, however, insofar as I am considering the possibility of algorithmic thought and how abstraction qua generalisation is a key operation in the respective ‘thinking’ structures of both humans and machines.<sup>17</sup> The key point is that these abstractive operations remain specific to the onto-epistemological grounds of humans, on the one hand, and machines, on the other – thus informing human modes of thought as well as algorithmic ones. For instance, returning to the case of the algorithmic recognition of handwritten zeros, the deep-learning model identifies and constructs representations that are more relevant than those that any human programmer could have identified and given to the machine. In fact, these are representations that a human would have not (and could have not) abstracted in the first place. The way the program extracts and organises information in terms of features and then generalises this information to form the desired ‘concept’ – or, in computational terms, the desired output representation of zero – is thus entirely and exclusively computational.<sup>18</sup>

We therefore must be careful when addressing how a human receives and elaborates stimuli or information, on the one hand, and how, on the other, a computing machine might also be said to do the same. It is important to talk here of *incommensurability* between the abstractive choices of humans and those of computing machines. ‘Incommensurability’ is the right word because the two cannot be measured against each other or compared by a common standard. Considering such an incommensurable dimension is particularly relevant in the context of debates about XAI because it allows us to highlight how explainability is a *representational problem* that pertains to *communication*. For abstractions to be successfully represented and thus expressed and shared, a common experience between the communicator and the receiver of the communication must be in place. Of course, this is not possible in the case of human-machine interactions, for no common phenomenological or existential ground exists between human abstractions and those of a computational agent. The specificity of computational abstraction and its suitability as the grounds of studying algorithmic thought are thus not claims strictly about cognitive science, as they do not emphasise the cognitive similarities and differences between abstractions by humans and machines. Rather, I stress the difficulty of comparing human and machine abstractive operations when ontological grounds shift and epistemic possibilities consequently vary. Acknowledging an

incommensurability between how humans and machines build models involves recognising this ontological and epistemological disparity between how humans and computational agents make decisions. Inevitably, this discrepancy is mirrored in how such decisions might be respectively recounted or represented by humans and artificial algorithmic agents.

Originating in ancient Greek mathematics, the notion of incommensurability denotes the absence of a common unit of measurement between two magnitudes. The development of this concept drove the distinction between geometry and arithmetic and is also central to the study of the ratios of numbers. Outside mathematics, however, the notion of incommensurability is used to denote that for which no shared nomenclature or shared ground for evaluation exists. In this sense, incommensurability is a key concept in 20th-century philosophy of science. Turning to this disciplinary context, in 1962 the philosophers of science Thomas Kuhn and Paul Feyerabend independently (but equally influentially) argued that successive scientific theories (with their associated concepts, methods and worldviews) are incommensurable.<sup>19</sup> For Feyerabend (1962), incommensurability was a semantic issue which he addressed to challenge conceptual conservatism in science and the approach to explanation, reduction and scientific advancement employed by logical positivism. In Kuhn's historical philosophy of science, too, incommensurability was a problem of language; for Kuhn (1962) as well, and to quote Michael Polanyi (whose philosophical work on the practice of science influenced both Kuhn and Feyerabend), scientists from diverse schools of thought and periods in time 'think differently, speak a different language, live in a different world' (Polanyi, 1958: 151). Beyond semantics, however, incommensurability was also a methodological and perceptual issue and a problem in taxonomy for Kuhn. He described as incommensurable the stark contrast between theoretical frameworks for which not only nomenclatures do not overlap but also for which no shared perceptions, methods or classifications exist.

In the philosophy of science, the notion of incommensurability is controversial; its meaning and usefulness are often contested, and discussions on this topic are never fully resolved.<sup>20</sup> I do not chronicle these discussions and their consequences here. Nonetheless, it is valuable to consider how the incommensurable has been introduced and addressed in that philosophical context and tradition of thought: this is because those debates help us situate incommensurability conceptually and, most importantly, because both Kuhn and Feyerabend proposed the concept while assessing the epistemic possibilities of scientific explanation. Precisely in relation to explanation, then, the notion of incommensurability confirms that explainability – in AI research as elsewhere – is a representational and communicational issue. Obviously, language plays a central role in this. Perhaps, to an extent, humans are bound to relate to

what they cannot represent or communicate with metaphors and analogies from their own experiences. So, for instance, we say that a computing machine ‘sees’, ‘listens’ or ‘thinks’, just as we say that an aeroplane ‘flies’ despite our awareness that an aircraft and a bird take flight in profoundly different ways.<sup>21</sup> In this respect, however, the challenge for both the philosophical and sociocultural studies of computational automation is to find or found the epistemological means to theorise, as well as possible, the incommensurable orders of intelligibility and sensibility that automated computational agents produce. Inevitably, the notion of incommensurability to be developed must transcend that proposed in the history of the philosophy of science: the long-term goal is not to apply Kuhn’s or Feyerabend’s respective understandings of the incommensurable to computational media and computational culture but to develop a radical version of the concept to address the specificities of human and algorithmic modes of abstraction.

### **‘Upon Opening the Black Box’**

To address this challenge, deep learning offers a particularly relevant case study. In the words of Yoshua Bengio, deep-learning research focuses on ‘learning algorithms that discover multiple levels of distributed representations, with higher levels representing more abstract concepts’ (2013: 1). ‘A deep learning algorithm’, Bengio continues, ‘is a particular kind of representation learning procedure that discovers *multiple levels of representation, with higher-level features representing more abstract aspects of the data*’ (2013: 2, emphasis in original). While much of computer programming has historically consisted in making human abstraction significant and operative within the instrumental remit of algorithmic machines, with deep learning we face the opposite case: the abstractions and consequent instructions the machine gives itself now require interpretation for them to be significant and operative for humans. The modes of organisation, categorisation and classification that belong to the abstractive operations of these computational cognitive agents are indeed incommensurable. Maintaining a theoretical focus on the nature and possibilities of abstraction as the balance moves between autonomy and automation within AI thus involves acknowledging and working with the prospect of modes of abstracting that might arise within calculation but also surpass the boundaries of human cognitive representation. In the example of recognising human handwriting, the ‘autonomy of automation’ (Fazi, 2019b: 94) regarding abstractive operations is demonstrated by a deep learning system producing internal representations independently from the phenomenological or experiential ground of the human programmer. Returning to this article’s opening example of AlphaGo Zero, such an autonomy is doubled: not only the outputs but also the training inputs are somewhat independent from human

knowledge. DeepMind's description of AlphaGo Zero as a form of superhuman intelligence is thus misleading; it would be more appropriate, from the point of view of incommensurability, to speak of *non-human* or *inhuman* intelligence (and the term 'intelligence' itself should also be problematised according to comparative epistemology).

Deep learning demonstrates that, when thinking and talking of computational cognitive agents, our theoretical efforts should attempt to move from strictly phenomenological analyses and existential qualifications (i.e. from efforts to address objects and situations as they appear to or are understood by human consciousness and through categories of human life and experience) towards more speculative modes of investigation. Adopting a speculative mode of investigation, we should address the critical prospect of understanding what explanation and interpretation might be in the formalising space of computation. Key to this speculative effort in relation to the study of computational automation is the possibility of constructing a theory of knowledge specific to computational artificial agents – a theory that can be advanced only by assessing the ontological and epistemological possibilities of machines. This theory would be valuable not only within the remit of digital studies but also for philosophical investigations of the relation between abstraction and experience and, consequently, the relation between rationality and the world. The following valuable programmatic point can then be drawn from the incommensurability debate in the philosophy of science. Both Kuhn's and Feyerabend's understandings of incommensurability have been accused of denying the possibility of progress and truth in science and thus implying irrationality.<sup>22</sup> This accusation, however, was rebutted by both scholars: claims about incommensurability do not imply that comparison is not possible but that it is much more difficult than the logical positivism and logical empiricism of the time assumed it was. Both Kuhn and Feyerabend made the notion of the incommensurable a powerful weapon in their post-positivist arsenals, and they used it to challenge evaluation and explanation based on absolute universal criteria or a neutral observation language.

Although differences certainly exist between the contexts and the aims of that debate – which pertained to the possibility of theory comparison – and the present study on deep learning and explainability, I propose that we can also mobilise the concept of incommensurability to problematise the 21st-century (implicit or explicit) positivist approaches to computational culture and society via data science.<sup>23</sup> Doing so does *not* imply relativism but, in fact, quite the opposite: I am arguing for the need to be loyal to the specificities of humans and machines in our comparisons. Similarly, we should not take for granted the fact that, when dealing with computing machines' abstractions that transcend the epistemic boundaries of human cognitive representation, we are working with models that are, at this time, *both within and beyond logos*. In other words, *these*

*models are logical* because they are computational and thus based on the possibility of a formal, logico-mathematical account of calculation; however, in a different sense, *they are also a-logical* because they are, at present, inexpressible or unrepresentable by humans (where ‘logos’, according to its ancient Greek etymology, not only means ‘reason’ or ‘proportion’ but also ‘word’, ‘discourse’, ‘speech’, and derives from the verb *légō*, ‘to count’, ‘to tell’, ‘to speak’). Focusing on the notion of incommensurability, then, allows us to emphasise the paradoxical condition of logico-mathematical abstraction in computation, which despite being a key tool for human attempts to organise and make sense of reality, today also surpasses that human-centric instrumental horizon with its AI implementations. For these AI-native models to be held rationally accountable, we should first ask whose and what rationality we are discussing. This question is radically open and acknowledges that a comparison between kinds and modes of thought is, to an extent, necessary to study AI’s ‘thinking’ procedures. The use of terms such as ‘thinking’ and ‘intelligence’ (which originated in human epistemology) does not contradict my argument; rather, their use confirms the inevitability of a comparison, although a comparison that will always be incomplete and partial because humans, as observers and interpreters, can only offer epistemic representations that have been shaped within their own ontological domain.

In this respect, it must be highlighted that incommensurability is a *translation failure*: on the one hand, a satisfactory translation between incommensurable entities is difficult or even impossible; on the other, a ‘translation failure’ also signals the limits of approaching explainable AI by searching for the quality or propriety of being translatable. It is important to stress this vis-à-vis current issues in the contemporary quest for fair, accountable, transparent AI precisely because that quest appears to be predicated on research that understands interpretability in terms of translation. It is thus also useful to consider how Kuhn (2000b) attributed the equation between translation and interpretation to the analytic tradition of philosophy. This equation was, in his opinion, misleading: incommensurability does not mean that a theoretical term, for instance, cannot be interpreted (that is, be made intelligible); rather, it means that it cannot be translated, as it has no equivalent in another theoretical language.

Returning to debates in and on XAI, such an equation between interpretation and translation can be observed in research that promotes the advancement of future XAI systems by developing new techniques able to produce *interpretable models* of machine-learning operations; these models, in turn, are paired with interfaces to advance useful *human-machine translations*, thus generating meaningful explanatory dialogues for end users. Interpretability via human-machine translation is, for instance, the explicit goal of the Defense Advanced Research Project Agency’s (DARPA) XAI initiative, which – recognising explainability




as a real issue for the computational systems of today and tomorrow – aims to develop human-centric perspectives in the design of artificial cognitive agents.<sup>24</sup> The challenge, for DARPA and other parties involved in the quest for interpretability in AI,<sup>25</sup> is to achieve understanding without compromising the predictive power and overall learning performance of the computational system. To do so, DARPA's XAI initiative encompasses various projects ranging from the design of entirely new kinds of deep neural networks comprising smaller and hence more easily understood modules to the borrowing of insights from the psychology of human expertise and decision-making.

What observations can be advanced about DARPA's XAI in relation to the issue of incommensurability? First, consistent with similar technoscientific attempts, DARPA's quest for XAI aims to bring what is beyond human knowledge back into the domain of human cognitive representation. Second, the goal of DARPA's XAI project is to find meaningful representations of the machine's own abstractions, even though these representations might be only useful or valuable to human actors and not, strictly speaking, necessary for the operativity of machine agency itself. Noticeably, there is not yet an obvious way of designing an artificial computational system that can explain itself, just as there is no consensus on what that explanation should look like – that is, what such an explanation should aim to represent. Third, then, the following question must be answered: would giving enough speculative credit and attention to the incommensurable operations of artificial cognitive systems be enough to produce such a shared account of a useful, successful explanation?

This question constitutes my conclusion; I leave it open to future research on the topic, which would have to further problematise the possibility of explanation in AI precisely because the opportunity of direct human-machine translations for artificial cognitive systems that are *de facto* beyond human representation can be questioned via the notion of incommensurability. In a famous polemical essay, Langdon Winner (1993) criticised imperatives of 'opening the black box' being obeyed, in his opinion, by the social construction of technology. Winner noted that, 'upon opening the black box', the risk was of 'finding it empty'. In a parallel yet distinct sense, we can borrow Winner's famous expression to consider now whether contemporary XAI's imperatives of opening the black box are running a similar risk. If there is indeed such a risk, it is less of finding the black box empty than of realising that there is nothing to translate or to render precisely because the possibility of human representation never existed in the first place.

## ORCID iD

M. Beatrice Fazi  <https://orcid.org/0000-0001-7183-8095>

## Notes

1. In Gilles Deleuze and Felix Guattari's *A Thousand Plateaus*, the two board games are compared philosophically. 'Chess', Deleuze and Guattari wrote, 'is a game of State'; its pieces 'have an internal nature and intrinsic properties from which their movements, situations, and confrontations derive.' The pieces in Go, in contrast, are 'pellets, disks, simple arithmetic units, and only have an anonymous, collective, third-person function'; they have 'no intrinsic properties, only situational ones' (2004: 389).
2. Alpaydin explains that reinforcement learning 'is also known as learning with a critic. The agent takes a sequence of actions and receives a reward/penalty only at the very end, with no feedback during the intermediate actions. Using this limited information, the agent should learn to generate the actions to maximize the reward in later trials' (2016: 180).
3. I use the term 'representation' not despite but because of debates about its crisis in science, art and philosophy (for an overview of some of these discussions, see Nöth, 2003). I do so to point to a renewed engagement with questions about the possibility (or impossibility) of representing. Moreover, I am speaking of representation because it 'lies at the heart of the debate between the logic-inspired and the neural-network-inspired paradigms for cognition' (LeCun et al., 2015: 441).
4. The first model of an artificial neuron was published by Warren McCulloch and Walter Pitts (1943), while the first general definition of machine learning was made by Arthur Samuel (1959). In 1957, Frank Rosenblatt (1962) developed the *perceptron*, an electronic device that implemented a simplified model of a biological neuron for pattern recognition. For an account of the development of deep learning, see Schmidhuber (2015).
5. In 1986, David Rumelhart, Geoffrey Hinton and Ronald Williams presented experimental evidence of the usefulness of the hidden dimension of artificial neural networks' back-propagation algorithms. In the journal *Nature*, they wrote: 'We describe a new learning procedure, back-propagation, for networks of neurone-like units. The procedure repeatedly adjusts the weights of the connections in the network so as to minimize a measure of the difference between the actual output vector of the net and the desired output vector. As a result of the weight adjustments, internal "hidden" units which are not part of the input or output come to represent important features of the task domain, and the regularities in the task are captured by the interactions of these units. The ability to create useful new features distinguishes back-propagation from earlier, simpler methods such as the perceptron-convergence procedure' (1986: 533; see also LeCun et al., 1998b).
6. For instance, parts of computer and data science stress that, at present, deep neural networks are often embedded within more traditional software, the main algorithmic architecture and strategy of which are known; it is also emphasised that the performance of deep learning can be at least partially understood at a theoretical level precisely because of this composite character of contemporary machine-learning systems. The AlphaGo Zero case itself exemplifies this point insofar as the deep-learning element of the program focuses on calculating two central functions of the system (the expert policy and the value approximation function). See Silver et al. (2017).

7. Whether deep learning needs a mathematical foundation is still debated; it is generally agreed, however, that mathematical justifications for deep learning's success remain elusive. New paradigms of mathematical reasoning are thus sought as well as new modes of analysis (see, for instance, the work of Sanjeev Arora at Princeton University).
8. In arguing this, I diverge from Bijker's claim (2010) that the technical and political question 'How to make technology?' can be answered by bracketing the philosophical question 'What is technology?'
9. The concept of agency is not antithetical to that of automation: something automated, such as a computational process, can have agency if we take the term to mean the capacity to produce a particular effect and understand automation as not synonymous with automatism.
10. Understanding and explanation are key concepts in epistemology and central topics in debates about scientific knowledge. Some views take understanding to be a psychological process involving the cognitive ability to explain; other positions instead argue that understanding is not necessarily explanatory (see Khalifa, 2017).
11. For explanatory and predictive modelling in statistics, see Shmueli (2010).
12. I am here hinting to the etymological origin of the term 'abstraction', which lies in the Latin verb *abstrahere*, meaning 'to draw away'.
13. This is true for the technoscientific literature on the topic but also for technocultural engagements in the field of media and software studies. Adrian Mackenzie and Anna Munster (2019), for example, have proposed the notion of 'platform seeing' to describe the computational operationalisation of a new mode of observing.
14. LeCun, Bottou, Bengio and Haffner (1998a) presented one of the most influential cases for the introduction of neural networks to recognise handwritten characters.
15. In supervised learning, algorithms process labelled datasets; while the inner relations of these data might be unknown, the needed output is known: 'the goal is to learn a function that maps from a set of input attributes for an instance to an accurate estimate of the missing value for the target attribute of the same instance' (Kelleher, 2019: 255). In unsupervised learning, instead, there is no target attribute or predefined output. 'The aim . . . is to find the regularities of the input' (Alpaydin, 2016: 111). The neural network attempts to find structure in the data by extracting useful features and analysing them.
16. See, for instance, research by Torralba and Freeman (2014).
17. In Geirhos et al. (2018), both human intelligence and machine intelligence are described as grounded in the power of generalisation that belongs equally to biological and artificial neural networks. In cognitive psychology, generalisation is understood as the basis of the process of learning from experience. In the literature on deep learning, generalisation is addressed in terms of the capacity of a model to learn from given data and then apply that information to other data.
18. Deep learning's artificial neurons respond to simple shapes and then more complex structures until they can address highly abstract concepts. Alpaydin explained that 'starting from the raw input, each hidden layer combines the values in its preceding layer and learns more complicated functions of the

- input. . . . Successive layers correspond to more abstract representations until we get to the final layer where the inputs are learned in terms of the most abstract concepts' (2016: 104). 'In deep learning,' Alpaydin continued, 'the idea is to learn feature levels of increasing abstraction with minimum human contribution . . . because in most applications, we do not know what structure there is in the input, especially as we go up, and the corresponding concepts become "hidden." . . . It is this extraction of hidden dependencies, or patterns, or regularities from data that allows abstraction and learning general descriptions' (2016: 106).
19. In this article, I can only briefly mention some differences between Kuhn's and Feyerabend's positions, which should not be conflated. The use and meaning of the concept of incommensurability also continued to evolve throughout both Kuhn's and Feyerabend's scholarship.
  20. For an overview of 'the incommensurability thesis', see Sankey (1994, 1997). Issues of scientific change and theory comparison are also addressed in Soler, Sankey and Hoyningen-Huene (2008).
  21. 'The quest for "artificial flight" succeeded when the Wright brothers and others stopped imitating birds and started . . . learning about aerodynamics' (Russell and Norvig, 2010: 3).
  22. 'My critics respond to my views on this subject with charges of irrationality, relativism, and the defense of mob rule. These are all labels which I categorically reject, even when they are used in my defense by Feyerabend. To say that, in matters of theory choice, the force of logic and observation cannot in principle be compelling is neither to discard logic and observation nor to suggest that they are not good reasons for favoring one theory over another' (Kuhn, 2000a: 126).
  23. I have discussed data science's positivist inclinations in Fazi (2017).
  24. The Defense Advanced Research Projects Agency (DARPA) is the US Department of Defense's body responsible for research and development projects in technology and science for use by the military.
  25. From a legislative perspective, it can be mentioned how the European Union (EU) has declared that EU citizens can challenge legal (or equally significant) decisions made by algorithms and appeal for human intervention and interpretation. This piece of legislation is part of the General Data Protection Regulation (GDPR) that went into effect in May 2018 and sketches the contours of a 'right to explanation'.

## References

- Alpaydin, Ethem (2016) *Machine Learning: The New AI*. Cambridge, MA: MIT Press.
- Amoore, Louise (2020) *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham, NC: Duke University Press.
- Ananny, Mike and Crawford, Kate (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20(3): 973–989.
- Beer, David (2018) *The Data Gaze: Capitalism, Power and Perception*. London: SAGE.
- Bengio, Yoshua (2013) Deep learning of representations: Looking forward. In: Dediu, Adrian-Horia, Martín-Vide, Carlos, Mitkov, Ruslan and Truthe,

- Bianca (eds) *Statistical Language and Speech Processing*. Berlin: Springer, pp. 1–37.
- Benjamin, Ruha (2019) *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity.
- Beyer, Mark A. and Laney, Douglas (2012) *The Importance of 'Big Data': A Definition*. Stamford, CT: Gartner Report.
- Bijker, Wiebe E. (2010) How is technology made? That is the question! *Cambridge Journal of Economics* 34(1): 63–76.
- Breiman, Leo (2001) Statistical modeling: The two cultures. *Statistical Science* 16(3): 199–231.
- Burrell, Jenna (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1): 1–12.
- Deleuze, Gilles and Guattari, Felix (2004) *A Thousand Plateaus: Capitalism and Schizophrenia*, trans. Massumi, Brian. London: Continuum.
- Domingos, Pedro (2015) *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. London: Allen Lane.
- Fazi, M. Beatrice (2017) The ends of media theory. *Media Theory* 1(1): 107–121.
- Fazi, M. Beatrice (2019a) Can a machine think (anything new)? Automation beyond simulation. *AI & Society* 34: 813–824.
- Fazi, M. Beatrice (2019b) Distraction machines? Augmentation, automation and attention in a computational age. *New Formations: A Journal of Culture, Theory, Politics* 98: 85–100.
- Feyerabend, Paul K. (1962) Explanation, reduction, and empiricism. In: Feigl, Herbert and Maxwell, Grover (eds) *Scientific Explanation, Space, and Time*. Minneapolis: University of Minneapolis Press, pp. 29–97.
- Geirhos, Robert, Medina Temme, Carlos R., Rauber, Jonas, Schütt, Heiko H., Bethge, Matthias and Wichmann, Felix A. (2018) Generalisation in humans and deep neural networks. In: Bengio, Samy, Wallach, Hanna M., Larochelle, Hugo, Grauman, Kristen, Cesa-Bianchi, Nicolò and Garnett, Roman (eds) *Advances in Neural Information Processing Systems 31 (NIPS 2018)*. Available at: <http://papers.nips.cc/paper/7982-generalisation-in-humans-and-deep-neural-networks.pdf> (accessed 7 August 2019).
- Gibney, Elizabeth (2017) Self-taught AI is best yet at strategy game Go. *Nature News*, 18 October. Available at: <https://www.nature.com/news/self-taught-ai-is-best-yet-at-strategy-game-go-1.22858> (accessed 7 August 2019).
- Hassabis, Demis and Silver, David (2017) AlphaGo Zero: Learning from scratch. *DeepMind.com* (blog). Available at: <https://deepmind.com/blog/alphago-zero-learning-scratch> (accessed 7 August 2019).
- Kelleher, John D. (2019) *Deep Learning*. Cambridge, MA: MIT Press.
- Khalifa, Kareem (2017) *Understanding, Explanation, and Scientific Knowledge*. Cambridge: Cambridge University Press.
- Kline, Ronald and Pinch, Trevor (1999) The social construction of technology. In MacKenzie, Donald and Wajcman, Judy (eds) *The Social Shaping of Technology (Second Edition)*. Maidenhead: Open University Press, pp. 113–115.
- Kuhn, Thomas S. (1962) *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Kuhn, Thomas S. (2000a) Reflections on my critics. In: Conant, James and Haugeland, John (eds) *The Road since Structure: Philosophical Essays*,

- 1970–1993, with an *Autobiographical Interview*. Chicago: University of Chicago Press, pp. 123–175.
- Kuhn, Thomas S. (2000b) Commensurability, comparability, communicability. In: Conant, James and Haugeland, John (eds) *The Road since Structure: Philosophical Essays, 1970–1993, with an Autobiographical Interview*. Chicago: University of Chicago Press, pp. 33–57.
- Latour, Bruno (1987) *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge, MA: Harvard University Press.
- Latour, Bruno (1999) *Pandora's Hope: Essays on the Reality of Science Studies*. Cambridge, MA: Harvard University Press.
- LeCun, Yann, Bengio, Yoshua and Hinton, Geoffrey (2015) Deep learning. *Nature* 521: 436–444.
- LeCun, Yann, Bottou, Leon, Bengio, Yoshua and Haffner, Patrick (1998a) Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11): 2278–2324.
- LeCun, Yann, Bottou, Leon, Orr, Genevieve B. and Müller, Klaus-Robert (1998b) Efficient backprop. In: Orr, Genevieve B. and Müller, Klaus-Robert (eds) *Neural Networks: Tricks of the Trade*. Berlin: Springer, pp. 9–50.
- Lin, Henry W., Tegmark, Max and Rolnick, David (2017) Why does deep and cheap learning work so well? *Journal of Statistical Physics* 168(6): 1223–1247.
- Mackenzie, Adrian (2015) The production of prediction: What does machine learning want? *European Journal of Cultural Studies* 18(4–5): 429–445.
- Mackenzie, Adrian (2018) *Machine Learners: Archaeology of a Data Practice*. Cambridge, MA: MIT Press.
- Mackenzie, Adrian and Munster, Anna (2019) Platform seeing: Image ensembles and their invisibilities. *Theory, Culture & Society* 36(5): 3–22.
- McCulloch, Warren S. and Pitts, Walter (1943) A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 5: 115–133.
- Morrison, Margaret (2015) *Reconstructing Reality: Models, Mathematics, and Simulations*. Oxford: Oxford University Press.
- Noble, Safiya Umoja (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.
- Nöth, Winfried (2003) Crisis of representation? *Semiotica* 143: 9–15.
- O'Neil, Cathy (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Penguin.
- Pasquale, Frank (2015) *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.
- Pinch, Trevor J. and Bijker, Wiebe E. (1984) The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science* 14(3): 399–441.
- Polanyi, Michael (1958) *Personal Knowledge: Towards a Post-Critical Philosophy*. London: Routledge & Kegan Paul.
- Rosenblatt, Frank (1962) *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, DC: Spartan.
- Rumelhart, David E., Hinton, Geoffrey E. and Williams, Ronald J. (1986) Learning representations by back-propagating errors. *Nature* 323: 533–536.
- Russell, Stuart J. and Norvig, Peter (2010) *Artificial Intelligence: A Modern Approach (Third Edition)*. Harlow: Pearson.

- Sample, Ian (2017) 'It's able to create knowledge itself': Google unveils AI that learns on its own. *The Guardian*, 18 October. Available at: <https://www.theguardian.com/science/2017/oct/18/its-able-to-create-knowledge-itself-google-unveils-ai-learns-all-on-its-own> (accessed 7 August 2019).
- Samuel, Arthur L. (1959) Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* 3(3): 210–229.
- Sankey, Howard (1994) *The Incommensurability Thesis*. London: Ashgate.
- Sankey, Howard (1997) Incommensurability: The current state of play. *Theoria: An International Journal for Theory, History and Foundations of Science* 12(3): 425–455.
- Schmidhuber, Juergen (2015) Deep learning in neural networks: An overview. *Neural Networks* 61: 85–117.
- Shmueli, Galit (2010) To explain or to predict? *Statistical Science* 25(3): 289–310.
- Silver, David, Schrittwieser, Julian, Simonyan, Karen, Antonoglou, Ioannis, Huang, Aja, Guez, Arthur, Hubert, Thomas et al. (2017) Mastering the game of Go without human knowledge. *Nature* 550: 354–359.
- Soler, Léna, Sankey, Howard and Hoyningen-Huene, Paul (eds) (2008) *Rethinking Scientific Change and Theory Comparison: Stabilities, Ruptures, Incommensurabilities?* Dordrecht: Springer.
- Spreuwenberg, Silvie (2019) *AIX: Artificial Intelligence Needs Explanation. Why and How Transparency Increases the Success of AI Solutions*. Amsterdam: LibRT BV.
- Torralba, Antonio and Freeman, William T. (2014) Accidental pinhole and pinspeck cameras: Revealing the scene outside the picture. *International Journal of Computer Vision* 110(2): 92–112.
- Turing, Alan M. (1936) On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* 42: 230–265.
- Turing, Alan M. (1950) Computing machinery and intelligence. *Mind* 59(236): 433–460.
- Winner, Langdon (1993) Upon opening the black box and finding it empty. *Science, Technology, & Human Values* 18(3): 362–378.

**M. Beatrice Fazi** is Lecturer in Digital Humanities in the School of Media, Arts and Humanities at the University of Sussex, United Kingdom. Her research focuses on the ontologies and epistemologies produced by contemporary technoscience, particularly in relation to issues in artificial intelligence and computation and to their impact on culture and society. She has published extensively on the limits and potentialities of the computational method, on digital aesthetics and on the automation of thought. Her monograph *Contingent Computation: Abstraction, Experience, and Indeterminacy in Computational Aesthetics* was published by Rowman & Littlefield International in 2018.

**This article is part of the *Theory, Culture & Society* special section on 'Algorithmic Thought' (TCS 38(7–8), December 2021), edited by M. Beatrice Fazi.**